1 Essentials

Matrix/Vector

Orthogonal: (i.e. columns are orthonormal!) $\mathbf{A}^{-1} = \mathbf{A}^\top$, $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$, $\det(\mathbf{A}) \in \{+1, -1\}$, $\det(\mathbf{A}^\top\mathbf{A}) = 1$

Inner Product: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top\mathbf{y} = \sum_{i=1}^N \mathbf{x}_i\mathbf{y}_i$. • $\langle \mathbf{x} \pm \mathbf{y}, \mathbf{x} \pm \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle \pm 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle$ • $\langle \mathbf{x}, \mathbf{y}+\mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ • $\langle \mathbf{x}+\mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$ • $\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \cdot \cos(\theta)$ • If $\mathbf{y}$ is a unit vector then $\langle \mathbf{x}, \mathbf{y} \rangle$ projects $\mathbf{x}$ onto $\mathbf{y}$

Outer Product: $\mathbf{u}\mathbf{v}^\top$, $(\mathbf{u}\mathbf{v}^\top)_{i,j} = \mathbf{u}_i\mathbf{v}_j$

Transpose: $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$

Cross product: $\vec{a} \times \vec{b} = (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)^\top$

Norms

$\|\mathbf{x}\|_0 = |\{i | x_i \neq 0\}|$ (non convex)

$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^N \mathbf{x}_i^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

$\|\mathbf{x}\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{\frac{1}{p}}$

$\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n \mathbf{m}_{i,j}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2} = \sqrt{trace(A\ A^H)}$

$\|\mathbf{M}\|_1 = \sum_{i,j} |m_{i,j}|$

$\|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M})$

$\|\mathbf{M}\|_p = \max_{\mathbf{v} \neq 0} \frac{\|\mathbf{M}\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$

$\|\mathbf{M}\|_\star = \sum_{i=1}^{\min(m,n)} \sigma_i$

Derivatives

$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$ $\quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$

$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}^\top + \mathbf{A})\mathbf{x}$ $\quad \frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x}) = \mathbf{A}^\top\mathbf{b}$

$\frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top$ $\quad \frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b}) = \mathbf{b}\mathbf{c}^\top$

$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$ $\quad \frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$

$\frac{\partial}{\partial \mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$

$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2) = 2(\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b})$

Eigenvalue / -vectors

Eigenvalue Problem: $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$

1. solve $\det(\mathbf{A} - \lambda\mathbf{I}) \overset{!}{=} 0$ resulting in $\{\lambda_i\}_i$
2. $\forall\lambda_i$: solve $(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{x}_i = \mathbf{0}$, for $\mathbf{x}_i$.

Eigendecomposition

$\mathbf{A} \in \mathbb{R}^{N \times N}$ then $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ with $\mathbf{Q} \in \mathbb{R}^{N \times N}$.

if fullrank: $\mathbf{A}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^{-1}$ and $(\mathbf{\Lambda}^{-1})_{i,i} = \frac{1}{\lambda_i}$.

if $\mathbf{A}$ symmetric: $A = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ ($\mathbf{Q}$ orthogonal).

Probability / Statistics

• $P(x) := Pr[X = x] := \sum_{y \in Y} P(x, y)$ • $P(x|y) := Pr[X = x | Y = y] := \frac{P(x,y)}{P(y)}$, if $P(y) > 0$ • $\forall y \in Y : \sum_{x \in X} P(x|y) = 1$ (property for any fixed y) • $P(x, y) = P(x|y)P(y)$ • $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$ (Bayes' rule) • $P(x|y) = P(x) \Leftrightarrow P(y|x) = P(y)$ (iff X, Y independent) • $P(x_1, \ldots, x_n) = \prod_{i=1}^n P(x_i)$ (iff IID)

2 Principle Component Analysis

$\mathbf{X} \in \mathbb{R}^{D \times N}$. N observations, K rank.

1. Empirical Mean: $\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^N \mathbf{x}_n$.
2. Center Data: $\overline{\mathbf{X}} = \mathbf{X} - [\overline{\mathbf{x}}, \ldots, \overline{\mathbf{x}}] = \mathbf{X} - \mathbf{M}$.
3. Cov.: $\Sigma = \frac{1}{N}\sum_{n=1}^N (\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^\top = \frac{1}{N}\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$.
4. Eigenvalue Decomposition: $\Sigma = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$.
5. Select $K < D$, only keep $\mathbf{U}_K, \boldsymbol{\lambda}_K$.
6. Transform data onto new Basis: $\overline{\mathbf{Z}}_K = \mathbf{U}_K^\top\overline{\mathbf{X}}$.
7. Reconstruct to original Basis: $\tilde{\overline{\mathbf{X}}} = \mathbf{U}_K\overline{\mathbf{Z}}_K$.
8. Reverse centering: $\tilde{\mathbf{X}} = \tilde{\overline{\mathbf{X}}} + \mathbf{M}$.

For compression save $\mathbf{U}_k, \overline{\mathbf{Z}}_K, \overline{\mathbf{x}}$.

$\mathbf{U}_k \in \mathbb{R}^{D \times K}, \Sigma \in \mathbb{R}^{D \times D}, \overline{\mathbf{Z}}_K \in \mathbb{R}^{K \times N}, \overline{\mathbf{X}} \in \mathbb{R}^{D \times N}$

Iterative View

Residual r_i: $x_i - \tilde{x}_i = (I - uu^T)x_i$

Cov of r: $\frac{1}{n}\sum_{i=1}^n (I - uu^T)x_i x_i^T (I - uu^T)^T = (I - uu^T)\Sigma(I - uu^T)^T = \Sigma - 2\Sigma uu^T + uu^T\Sigma uu^T = \Sigma - \lambda uu^T$?

1. Find principal eigenvector of $(\Sigma - \lambda uu^T)$
2. which is the second eigenvector of Σ
3. iterating to get d principal eigenvector of Σ

Power Method

Power iteration: $v_{t+1} = \frac{Av_t}{\|Av_t\|}$, $\lim_{t \to \infty} v_t = u_1$

Assuming $\langle u_1, v_0 \rangle \neq 0$ and $|\lambda_1| > |\lambda_j| (\forall j \geq 2)$

3 SVD

$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})} d_{k,k} u_k (v_k)^\top$

$\mathbf{A} \in \mathbb{R}^{N \times P}, \mathbf{U} \in \mathbb{R}^{N \times N}, \mathbf{D} \in \mathbb{R}^{N \times P}, \mathbf{V} \in \mathbb{R}^{P \times P}$

$\mathbf{U}^\top\mathbf{U} = I = \mathbf{V}^\top\mathbf{V}$ ($\mathbf{U}$, $\mathbf{V}$orthonormal)

$\mathbf{U}$ columns are eigenvectors of $\mathbf{A}\mathbf{A}^\top$, $\mathbf{V}$ columns are eigenvectors of $\mathbf{A}^\top\mathbf{A}$, $\mathbf{D}$ diagonal elements are singular values.

$(\mathbf{D}^{-1})_{i,i} = \frac{1}{\mathbf{D}_{i,i}}$ (don't forget to transpose)

1. calculate $\mathbf{A}^\top\mathbf{A}$.
2. calculate eigenvalues of $\mathbf{A}^\top\mathbf{A}$, the square root of them, in descending order, are the diagonal elements of $\mathbf{D}$.
3. calculate eigenvectors of $\mathbf{A}^\top\mathbf{A}$ using the eigenvalues resulting in the columns of $\mathbf{V}$.
4. calculate the missing matrix: $\mathbf{U} = \mathbf{A}\mathbf{V}\mathbf{D}^{-1}$.
5. normalize each column of $\mathbf{U}$ and $\mathbf{V}$.

Low-Rank approximation

Using only K largest eigenvalues and corresponding eigenvectors. $\tilde{\mathbf{A}}_{i,j} = \sum_{k=1}^K \mathbf{U}_{i,k}\mathbf{D}_{k,k}\mathbf{V}_{j,k} = \sum_{k=1}^K \mathbf{U}_{i,k}\mathbf{D}_{k,k}(\mathbf{V}^\top)_{k,j}$.

Echart-Young Theorem

$\min_{rank(B)=K} \|A - B\|_F^2 = \|A - A_k\|_F^2 = \sum_{r=k+1}^{rank(A)} \sigma_r^2$

$\min_{rank(B)=K} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}$

4 Optimization

Alternating Least Squares

$f(U, v_i) = \sum_{(i,j) \in I} (a_{i,j} - \langle u_j, v_i \rangle)^2$

$f(u_i, V) = \sum_{(i,j) \in I} (a_{i,j} - \langle u_j, v_i \rangle)^2$

Least squares problems which are convex.

Coordinate Descent

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$
2. for $t = 0$ to $maxIter$:
3. sample $d \in_{u.a.r.} \{1, \ldots, D\}$
4. $u^\star = \arg\min_{u \in \mathbb{R}} f(x_1^{(t)}, .., x_{d-1}^{(t)}, u, x_{d+1}^{(t)}, .., x_D^{(t)})$
5. $\mathbf{x}_d^{(t+1)} = u^\star$ and $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)}$ for $i \neq d$

Gradient Descent (or Deepest Descent)

Gradient: $\nabla f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D}\right)^\top$

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$
2. for $t = 0$ to $maxIter$:
3. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})$, usually $\gamma \approx \frac{1}{t}$

Stochastic Gradient Descent (SGD)

Assume **Additive Objective:** $f(x) = \frac{1}{N}\sum_{n=1}^N f_n(x)$

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$
2. for $t = 0$ to $maxIter$:
3. sample $n \in_{u.a.r.} \{1, \ldots, N\}$
4. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f_n(\mathbf{x}^{(t)})$, typically $\gamma \approx \frac{1}{t}$.

Projected Gradient Descent (Constrained Opt.)

minimize $f(x)$, $x \in Q$ (constraint).

Project x onto Q: $P_Q(\mathbf{x}) = \arg\min_{y \in Q} \|\mathbf{y} - \mathbf{x}\|$,

Update: $\mathbf{x}^{(t+1)} = P_Q[\mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})]$,

$\mathbf{x}^{(t+1)}$ is unique if Q convex.

Lagrangian Multipliers

Minimize $f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) \leq 0$, $i = 1, .., m$ (**inequality constr.**) and $h_i(\mathbf{x}) = \mathbf{a}_i^\top\mathbf{x} - b_i = 0$, $i = 1, .., p$ (**equality constraint**)

Lagrangian: $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) := f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$

Dual function: $D(\boldsymbol{\lambda}, \boldsymbol{\nu}) := \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \in \mathbb{R}$

Dual Problem: $\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} D(\boldsymbol{\lambda}, \boldsymbol{\nu})$ s.t. $\boldsymbol{\lambda} \geq \mathbf{0}$. Note: $\max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} D(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq \min_{\mathbf{x}} f(\mathbf{x})$, equality if $dom(f)$ and f convex

Convex Optimization

Def.: $\{(x, t) | x \in dom(f), f(x) \leq t\}$, $f : \mathbb{R}^D \to \mathbb{R}$ is convex, if $dom(f)$ is a convex set, and if $\forall \mathbf{x}, \mathbf{y} \in dom(f)$, and for $0 \leq \alpha \leq 1$: $f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha)f(\mathbf{y})$. local=global min, **Convergence:** $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{c}{t}$. **Subgradient** $g \in \mathbb{R}^D$ of f at $\mathbf{x}$: $f(\mathbf{y}) \geq f(\mathbf{x}) + g^\top(\mathbf{y} - \mathbf{x})\ \forall \mathbf{y}$

Convex Relaxation

Replace non-convex rank constraints by convex norm constraints.

Nuclear norm: $\|A\|_* = \sum_i \sigma_i$

5 Non-Negative Matrix Factorization

Context Model: $p(w|d) = \sum_{z=1}^K p(w|z)p(z|d)$

Conditional independence assumption (*): $p(w|d) = \sum_z p(w, z|d) = \sum_z p(w|d, z)p(z|d) \overset{*}{=} \sum_z p(w|z)p(z|d)$

Symmetric parameterization: $p(w, d) = \sum_z p(z)p(w|z)p(d|z)$

EM for pLSA

Log-Likelihood: $L(\mathbf{U}, \mathbf{V}) = \sum_{i,j} x_{i,j} \log p(w_j|d_i)$

$= \sum_{(i,j) \in \mathbf{X}} \log \sum_{z=1}^K p(w_j|z)p(z|d_i)$

$p(w_j|z) = v_{zj}$, $p(z|d_i) = u_{zi}$, $\sum_j^N v_{zj} = \sum_z^K u_{zi} = 1$

E-Step (optimal q):

$q_{zij} = \frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^K p(w_j|k)p(k|d_i)} := \frac{v_{zj}u_{zi}}{\sum_{k=1}^K v_{kj}u_{ki}}$

M-Steps:

$p(z|d_i) = \frac{\sum_j x_{ij}q_{zij}}{\sum_j x_{ij}}$, $p(w_j|z) = \frac{\sum_i x_{ij}q_{zij}}{\sum_{i,l} x_{il}q_{zil}}$

Latent Dirichlet Allocation

To sample a new document, we need to extend X and U^T with a new row, s.t. $X = U^T V$. pLSA fixes both dimensions.

Dirichlet distribution: $p(u_i|\alpha) \propto \prod_{z=1}^K u_{zi}^{\alpha_z - 1}$

LDA model: $p(x|V, u) = \frac{l!}{\prod_j x_j!} \prod_j \pi_j^{x_j}$

where $\pi_j = \sum_z v_{zj}u_z$, $l = \sum_j x_j$

NMF Algorithm for quadratic cost function

$\mathbf{X} \in \mathbb{Z}_{\geq 0}^{N \times M}$, NMF: $\mathbf{X} \approx \mathbf{U}^\top\mathbf{V}$,

$\min_{\mathbf{U},\mathbf{V}} J(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2$

s.t. $\forall i, j, z : u_{zi}, v_{zj} \geq 0$

1. init: $\mathbf{U}, \mathbf{V} = rand()$ 2. repeat for $maxIters$:
3. upd. $(\mathbf{V}\mathbf{V}^\top)\mathbf{U} = \mathbf{V}\mathbf{X}^\top$, proj. $u_{zi} = \max\{0, u_{zi}\}$
4. update $(\mathbf{U}\mathbf{U}^\top)\mathbf{V} = \mathbf{U}\mathbf{X}$, proj. $v_{zj} = \max\{0, v_{zj}\}$

6 Word Embeddings

Distributional Model: $p_\theta(w|w') = \Pr[w \text{ occurs in context of } w']$

Log-likelihood: $L(\theta; \mathbf{w}) = \sum_{t=1}^T \sum_{\Delta \in I} \log p_\theta(w^{(t+\Delta)}|w^{(t)})$

Latent Vector Model: $w \to (\mathbf{x}_w, b_w) \in \mathbb{R}^{D+1}$

$p_\theta(w|w') = \frac{\exp[\langle \mathbf{x}_w, \mathbf{x}_{w'} \rangle + b_w]}{\sum_{v \in V} \exp[\langle \mathbf{x}_v, \mathbf{x}_{w'} \rangle + b_v]}$ (soft-max).

Modifications: $\log p_\theta(w|w') = \langle y_w, x_{w'} \rangle + b_w$, word y_w, c'txt $x_{w'}$

use GloVe objective

negative sampling (logistic classification)

GloVe (Weighted Square Loss)

Co-occurence Matrix: $\mathbf{N} = (n_{ij}) \in \mathbb{R}^{|V| \times |C|} = \#of word w_i \text{ in context } w_j$

Objective: $H(\theta; \mathbf{N})$

$= \sum_{n_{ij} > 0} f(n_{ij})(\log n_{ij} - \log \exp[\langle \mathbf{x}_i, \mathbf{y}_j \rangle + b_i + d_j])^2$

with $f(n) = \min\{1, (\frac{n}{n_{max}})^\alpha\}$, $\alpha \in (0; 1]$.

unnormalized distr. → 2-sided loss function
1. sample $(i,j) u.a.r, s.t. n_{ij} > 0$
2. $\mathbf{x}_i^{new} \leftarrow \mathbf{x}_i + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle)\mathbf{y}_j$
3. $\mathbf{y}_j^{new} \leftarrow \mathbf{y}_j + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle)\mathbf{x}_i$

Discussion

Word embeddings can model analogies and relatedness, but antonyms are usually not well captured.

7 K-means Algorithm

Target: $\min_{\mathbf{U},\mathbf{Z}} J(\mathbf{U},\mathbf{Z}) = \|\mathbf{X} - \mathbf{UZ}\|_F^2$
$= \sum_{n=1}^N \sum_{k=1}^K \mathbf{z}_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2$

1. **Initiate:** choose K centroids $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$
2. **Assign:** data points to clusters. $k^\star(\mathbf{x}_n) = \arg\min_k \{\|\mathbf{x}_n - \mathbf{u}_k\|_2\}$ returns cluster $k^\star$, whose centroid $\mathbf{u}_{k^\star}$ is closest to data point $\mathbf{x}_n$. Set $\mathbf{z}_{k^\star,n} = 1$, and for $l \neq k^\star$ $\mathbf{z}_{l,n} = 0$.
3. **Update** centroids: $\mathbf{u}_k = \frac{\sum_{n=1}^N z_{k,n}\mathbf{x}_n}{\sum_{n=1}^N z_{k,n}}$.
4. Repeat from step 2, stops if $\|\mathbf{Z} - \mathbf{Z}^{\text{new}}\|_0 = \|\mathbf{Z} - \mathbf{Z}^{\text{new}}\|_F^2 = 0$.

Computational cost: $O(k \cdot n \cdot d)$

8 Gaussian Mixture Models (GMM)

For GMM let $\theta_k = (\mu_k, \Sigma_k)$; $p_{\theta_k}(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$

Mixture Models: $p_\theta(\mathbf{x}) = \sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x})$

Assignment variable (generative model):
$z_{ij} \in [0,1]$, $\sum_{j=1}^k z_{ij} = 1$
$\Pr(z_k = 1) = \pi_k \Leftrightarrow p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$

Complete data distribution:
$p_\theta(\mathbf{x},\mathbf{z}) = \prod_{k=1}^K \left(\pi_k p_{\theta_k}(\mathbf{x})\right)^{z_k}$

Posterior Probabilities:
$\Pr(z_k = 1|\mathbf{x}) \overset{\text{Bayes}}{=} \frac{\Pr(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{l=1}^K \Pr(z_l=1)p(\mathbf{x}|z_l=1)} = \frac{\pi_k p_{\theta_k}(\mathbf{x})}{\sum_{l=1}^K \pi_l p_{\theta_l}(\mathbf{x})}$

Likelihood of observed data X:
$p_\theta(\mathbf{X}) = \prod_{n=1}^N p_\theta(\mathbf{x}_n) = \prod_{n=1}^N \left(\sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x}_n)\right)$

Max. Likelihood Estimation (MLE):
$\arg\max_\theta \sum_{n=1}^N \log\left(\sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x}_n)\right)$
$\geq \sum_{n=1}^N \sum_{k=1}^K q_k [\log p_{\theta_k}(\mathbf{x}_n) + \log \pi_k - \log q_k]$
with $\sum_{k=1}^K q_k = 1$ by Jensen Inequality.

Generative Model

1. sample cluster index $j \sim Categorical(\pi)$
2. given j, sample data $x \sim \text{Normal}(\mu_j, \Sigma_j)$

Expectation-Maximization (EM) for GMM

E-Step:
$\Pr[z_{k,n} = 1|\mathbf{x}_n] = q_{k,n} = \frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{x}_n|\mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \mathcal{N}(\mathbf{x}_n|\mu_j^{(t-1)}, \Sigma_j^{(t-1)})}$

M-Step: $\mu_k^{(t)} := \frac{\sum_{n=1}^N q_{k,n}\mathbf{x}_n}{\sum_{n=1}^N q_{k,n}}$, $\pi_k^{(t)} := \frac{1}{N}\sum_{n=1}^N q_{k,n}$

$\Sigma_k^{(t)} = \frac{\sum_{n=1}^N q_{k,n}(\mathbf{x}_n - \mu_k^{(t)})(\mathbf{x}_n - \mu_k^{(t)})^\top}{\sum_{n=1}^N q_{k,n}}$

Gaussian distribution

Standard deviation σ, Mean μ

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

Covariance matrix Σ, Mean μ

$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

Discussion K-means vs. EM

hard assignment vs soft. spherical clusters shapes vs covariance matrix. fast vs slow and more iteration. K-means can be used as initialization for EM.

K-means is a special case of GMM with covariances $\Sigma_j = \sigma^2 * I$. in the limit of $\sigma \to 0$, recover K-means.

Model Order Selection (AIC / BIC for GMM)

Trade-off between data fit (i.e. likelihood $p(\mathbf{X}|\theta)$) and complexity (i.e. # of free parameters $\kappa(\cdot)$). For choosing K:

Akaike Information Criterion: $\text{AIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X}) + \kappa(\theta)$

Bayesian Information Criterion: $\text{BIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X}) + \frac{1}{2}\kappa(\theta)\log N$

# of free params, fixed covariance matrix: $\kappa(\theta) = K \cdot D + (K-1)$ (K: # clusters, D: dim(data) = dim(μ_i), $K-1$: # free clusters, full covariance matrix: $\kappa(\theta) = K(D + \frac{D(D+1)}{2}) + (K-1)$.

Compare AIC/BIC for different K – the smaller the better. BIC penalizes complexity more.

9 Convolutional Neural Networks

Neurons: $F_\sigma(\mathbf{x};\mathbf{w}) = \sigma(w_0 + \sum_{i=1}^M x_i w_i)$.

Activation: tanh, ReLu, $s(x) = \frac{1}{1+e^{-x}}$, $s'(x) = s(x)(1-s(x))$

Output: linear regression: $\hat{\mathbf{y}} = \mathbf{W}^L \mathbf{x}^{L-1}$

binary classification (logistic):
$\hat{y}_1 = \hat{\mathrm{P}}[Y=1|\mathbf{x}] = \frac{1}{1+\exp[-\langle \mathbf{w}_1^L, \mathbf{x}^{L-1}\rangle]}$

multiclass (soft-max):
$\hat{y}_k = \hat{\mathrm{P}}[Y=k|\mathbf{x}] = \frac{\exp[\langle \mathbf{w}_k^L, \mathbf{x}^{L-1}\rangle]}{\sum_{m=1}^K \exp[\langle \mathbf{w}_m^L, \mathbf{x}^{L-1}\rangle]}$.

Loss function squared loss: $\frac{1}{2}(y-\hat{y})^2$

cross-entropy loss: $-y\log\hat{y} - (1-y)\log(1-\hat{y})$.

Units and Layers: layer-to-layer fwd. prop. notation: $\mathbf{x}^l = \sigma^l(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)})$. L-layer network:
$\mathbf{y} = \sigma^{(L)}\left(\mathbf{W}^{(L)}\sigma^{(L-1)}\left(\cdots\left(\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x}\right)\cdots\right)\right)\right)$

Backpropagation

Layer-to-layer Jacobian: $\mathbf{x}$ = prev. layer activation, $\mathbf{x}^+$ = next layer activation. Jacobian matrix $\mathbf{J} = J_{ij}$ of mapping $\mathbf{x} \to \mathbf{x}^+$, $\mathbf{x}_i^+ = \sigma(\mathbf{w}_i^\top \mathbf{x})$,

$J_{ij} = \frac{\partial \mathbf{x}_i^+}{\partial \mathbf{x}_j} = w_{ij} \cdot \sigma'(\mathbf{w}_i^\top \mathbf{x})$.

Across multiple layers: $\frac{\partial \mathbf{x}^{(l)}}{\partial \mathbf{x}^{(l-n)}} = \mathbf{J}^{(l)} \cdot \frac{\partial \mathbf{x}^{(l-1)}}{\partial \mathbf{x}^{(l-n)}} = \mathbf{J}^{(l)} \cdot \mathbf{J}^{(l-1)} \ldots \mathbf{J}^{(l-n+1)}$ and then back prop. $\nabla_{x^{(l)}}^\top \ell = \nabla_y^\top \ell \cdot \mathbf{J}^{(L)} \ldots \mathbf{J}^{(l+1)}$

Neural Networks for Images

Translation invariance of images → neurons compute same fct, shift invariant filters; weights defined as filter masks, e.g. convolution: $F_{n,m}(\mathbf{x};\mathbf{w}) = \sigma(b + \sum_{k=-2}^2 \sum_{l=-2}^2 w_{k,l} x_{n+k,m+l})$.

To reduce dimension of convolution, use {max, avg}-pooling

10 Sparse Coding

Orthogonal Basis

For $\mathbf{x}$ and o.n.b. $\mathbf{U}$ compute $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$. Approx $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$, $\hat{z}_i = z_i$ if $|z_i| > \epsilon$ else 0. Reconstruction Error $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{d \notin \sigma} \langle \mathbf{x}, \mathbf{u}_d \rangle^2$. Choice of base depends on signal. Fourier for global, wavelet for local support. PCA basis optimal for given Σ.

Overcomplete Basis

$\mathbf{U} \in \mathbb{R}^{D \times L}$ for # atoms = $L > D$ = dim(data). Decoding involved → add constraint $\mathbf{z}^\star \in \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\mathbf{x} = \mathbf{U}\mathbf{z}$. NP-hard → approximate with 1-norm (convex) or with MP.

Coherence • $m(\mathbf{U}) = \max_{i,j:i\neq j} |\mathbf{u}_i^\top \mathbf{u}_j|$ • $m(\mathbf{B}) = 0$ if $\mathbf{B}$ orthogonal matrix • $m([\mathbf{B},\mathbf{u}]) \geq \frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to orthogonal basis $\mathbf{B}$ (o.n.b. = orthonormal base)

Matching Pursuit (MP) approximation of $\mathbf{x}$ onto $\mathbf{U}$, using K entries. Objective: $\mathbf{z}^\star \in \arg\min_{\mathbf{z}} \|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2$, s.t. $\|\mathbf{z}\|_0 \leq K$ 1. init: $z \leftarrow 0, r \leftarrow x$ 2. while $\|\mathbf{z}\|_0 < K$ do 3. select atom with smallest angle $i^\star = \arg\max_i |\langle \mathbf{u}_i, \mathbf{r}\rangle|$ 4. update coefficients: $z_{i^\star} \leftarrow z_{i^\star} + \langle \mathbf{u}_{i^\star}, \mathbf{r}\rangle$ 5. update residual: $\mathbf{r} \leftarrow \mathbf{r} - \langle \mathbf{u}_{i^\star}, \mathbf{r}\rangle \mathbf{u}_{i^\star}$.

Exact recovery when: $K < 1/2(1 + 1/m(\mathbf{U}))$

Compressive Sensing: Compress data while gathering: • $\mathbf{x} \in \mathbb{R}^D$, K-sparse in o.n.b. $\mathbf{U}$ • $\mathbf{y} \in \mathbb{R}^M$ with $y_i = \langle \mathbf{w}_i, \mathbf{x}\rangle$: M lin. combinations of signal; $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} = \theta\mathbf{z}$, $\theta \in \mathbb{R}^{M\times D}$ • Reconstruct $\mathbf{x} \in \mathbb{R}^D$ from $\mathbf{y}$; find $\mathbf{z}^\star \in \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0$, s.t. $\mathbf{y} = \theta\mathbf{z}$ (e.g. with MP). Given $\mathbf{z}$, reconstruct $\mathbf{x}$ via $\mathbf{x} = \mathbf{U}\mathbf{z}$

Sufficient conditions: • $\mathbf{W}$ = Gaussian random projection, i.e. $w_{ij} \sim \mathcal{N}(0, \frac{1}{D})$ • $M \geq cK\log(\frac{D}{K})$, where c is some constant

Dictionary Learning

Adapt the dictionary to signal characteristics. Objective: $(\mathbf{U}^\star, \mathbf{Z}^\star) \in \arg\min_{\mathbf{U},\mathbf{Z}} \|\mathbf{X} - \mathbf{U}\cdot\mathbf{Z}\|_F^2$ not jointly convex but convex in 1 argument.

Matrix Factorization by Iter Greedy Minimization 1. Coding step: $\mathbf{Z}^{t+1} \in \arg\min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{U}^t\mathbf{Z}\|_F^2$ subject to $\mathbf{Z}$ being sparse ($\mathbf{z}_n^{t+1} \in \arg\min_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\|\mathbf{x}_n - \mathbf{U}^t\mathbf{z}\|_2 \leq \sigma\|\mathbf{x}_n\|_2$)
2. Dict update step: $\mathbf{U}^{t+1} \in \arg\min_{\mathbf{U}} \|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2$, subj to $\forall l \in [L] : \|\mathbf{u}_l\|_2 = 1$. (set $\mathbf{U} = [\mathbf{u}_1^t \cdots \mathbf{u}_l \cdots \mathbf{u}_L^t]$, $\min_{\mathbf{u}_l} \|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2 = \min_{\mathbf{u}_l} \|\mathbf{R}_l^t - \mathbf{u}_l(\mathbf{z}_l^{t+1})^\top\|_F^2$ with $\mathbf{R}_l^t = \tilde{\mathbf{U}}\tilde{\Sigma}\tilde{\mathbf{V}}^\top$ by $\mathbf{u}_l^* = \tilde{\mathbf{u}}_1$)

11 Robust PCA

Idea: Approximate $\mathbf{X}$ with $\mathbf{L} \oplus \mathbf{S}$, $\mathbf{L}$ is low-rank, $\mathbf{S}$ is sparse.

$\min_{\mathbf{L},\mathbf{S}} \text{rank}(\mathbf{L}) + \mu\|\mathbf{S}\|_0$, s. t. $\mathbf{L} + \mathbf{S} = \mathbf{X}$. As non-convex, change to $\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1$ (*not* the same in general) Deal w/ missing values: subj. to $L_{ij} + S_{ij} = X_{ij}, \forall (i,j) \in \Omega_{observed}$

Perfect reconstruction is *not* possible if $\mathbf{S}$ is low-rank, $\mathbf{L}$ is sparse, or $\mathbf{X}$ is low-rank *and* sparse. Formally coherence: $\|\mathbf{U}^\top \mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{V}^\top \mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{U}\mathbf{V}^\top\|_{ij}^2 \leq \frac{\nu r}{n^2}$: $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$

Dual Ascent (Gradient Method for Dual Problem)

$\lambda^{t+1} = \lambda^t + \eta \nabla D(\lambda^t)$, $\nabla D(\lambda) = \mathbf{A}\mathbf{x}^* - \mathbf{b}$ for $\mathbf{x}^* \in \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x},\lambda)$

Dual Decomposition for Dual Ascent:

$\mathbf{x}_i^{t+1} := \arg\min_{\mathbf{x}_i} \mathcal{L}_i(\mathbf{x}_i, \lambda^t)$; $\lambda^{t+1} := \lambda^t + \eta^t\left(\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{t+1} - \mathbf{b}\right)$

Alternating Direction Method of Multipliers (ADMM)

$\min_{\mathbf{x}_1,\mathbf{x}_2} f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$ s. t. $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$, f_1, f_2 convex

Augmented Lagrangian:

$L_p(\mathbf{x}_1,\mathbf{x}_2,\boldsymbol{\nu}) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \boldsymbol{\nu}^\top(\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}) + \frac{p}{2}\|\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}\|_2^2$

ADMM:

$\mathbf{x}_1^{(t+1)} := \arg\min_{\mathbf{x}_1} L_p(\mathbf{x}_1, \mathbf{x}_2^{(t)}, \boldsymbol{\nu}^{(t)})$

$\mathbf{x}_2^{(t+1)} := \arg\min_{\mathbf{x}_2} L_p(\mathbf{x}_1^{(t+1)}, \mathbf{x}_2, \boldsymbol{\nu}^{(t)})$

$\boldsymbol{\nu}^{(t+1)} := \boldsymbol{\nu}^{(t)} + p(\mathbf{A}_1\mathbf{x}_1^{(t+1)} + \mathbf{A}_2\mathbf{x}_2^{(t+1)} - \mathbf{b})$

ADMM for RPCA:

$f_1(\mathbf{L}) = \|\mathbf{L}\|_\star$, $f_2(\mathbf{S}) = \lambda\|\mathbf{S}\|_1$, $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$ becomes $\mathbf{L} + \mathbf{S} = \mathbf{X}$, therefore $L_p(\mathbf{L},\mathbf{S},\boldsymbol{\nu}) = \|\mathbf{L}\|_\star + \nu\|\mathbf{S}\|_1 + \langle \boldsymbol{\nu}, \text{vec}(\mathbf{L} + \mathbf{S} - \mathbf{X})\rangle + \frac{p}{2}\|\mathbf{L} + \mathbf{S} - \mathbf{X}\|_F^2$